



Finding SNPs in a Sequence Using BLAST

Exploring the Sequence Viewer display of BLAST results for SNP discovery

<https://blast.ncbi.nlm.nih.gov> & <https://www.ncbi.nlm.nih.gov/snp/>

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Background

NCBI maintains a large public repository for human Short Nucleotide Variations database, also known as dbSNP. It can be accessed through the web from its homepage at www.ncbi.nlm.nih.gov/snp. These variation data are fully integrated with other NCBI sequence datasets, as well as its flagship sequence alignment tool - BLAST sequence alignment service at blast.ncbi.nlm.nih.gov.



NCBI used to provide a BLAST service to search against the flanking sequences of existing SNVs. Using this service in variation identification was cumbersome for several reasons:

- Variations identified using new technologies do not provide their true flanking sequences
- Variations identified from array experiments have very short flanking sequences, some of which could map to multiple genomic locations
- Long flanking sequences for some variations often have ambiguous base calls making variant calls and their mapping difficult
- Insertion and deletion alleles are generally represented by an N at the FASTA sequence level, which makes BLAST searches and their interpretation difficult without referencing the allele in the FASTA definition line
- Lastly, this approach did not provide the important genomic context and coordinates of identified variations

However, many reported human nucleotide variations in the biomedical literature often uses allele plus some flanking sequences to represent the variations, this makes correlating the all important phenotype and disease impact information to existing rsIDs in dbSNP as well as examine them in the proper genomic context difficult to achieve.

In this tutorial, we will demonstrate a way to map variants, contained within example sequences, using NCBI BLAST tool, through examining the genomic or mRNA alignment through the Graphical Sequence Viewer (SV) [1] linked embedded in the BLAST report [2], as well as interactive manipulating the display to better present the results in the context of gene features and SNV annotation.

Practical applications

Researchers often attempt to identify variations in specific genes or regions of the genome. *In silico* analysis of specific sequences by alignment to those in public collections can identify patterns of mismatches, which could represent a potential variation. Mapping variations to existing records in dbSNP can help connect them to functional analyses reported in literature and facilitate decision-making processes to enhance the speed and productivity of research. For example, two query sequences, a genomic sequence fragment [GQ892012.1](https://www.ncbi.nlm.nih.gov/nuccore/GQ892012.1) and a cDNA clone [CK130262.1](https://www.ncbi.nlm.nih.gov/nuccore/CK130262.1), are related to the human IL4 gene. According to published articles (go.usa.gov/xGXRK), they may contain variations that have been shown to affect the disease prognosis. In the following sections of this handout, we will demonstrate how to identify sequence variations contained in these two sequence records through BLAST alignment and match them to highly informative records in dbSNP.

Setting up the BLAST searches

Selection of BLAST algorithm

To identify dbSNP records that map to the query sequence, select “nucleotide blast” from the “Web BLAST” section on the BLAST homepage (blast.ncbi.nlm.nih.gov).

Query input

The BLAST search pages support queries in raw sequence, sequences in FASTA format, or identifiers of sequence (GIs or accessions). Accession numbers [GQ892012.1](https://www.ncbi.nlm.nih.gov/nuccore/GQ892012.1) or [CK130262.1](https://www.ncbi.nlm.nih.gov/nuccore/CK130262.1) will be used in the Query box in this set of exercises.

Database setup

Different types of input sequences require different databases. With genomic DNA sequences as the input query, select “RefSeq Representative genomes” as the target database. To search with expressed sequences, select “Reference RNA sequences (refseq_rna)” instead. You can further restrict database lookup to a particular organism using the Organism input box. This limit speeds up the search, makes the results returned more focused, and greatly simplifies the subsequent analyses. Submit the search by clicking the “BLAST” button. The web page automatically checks for results during the search, and renders the results in the browser window when the search is completed.

For your convenience, pre-configured BLAST search pages for a human cDNA and genomic query are listed below: go.usa.gov/xGX9Q for the genomic query (GQ892012.1), and go.usa.gov/xGXMn for the cDNA query (CK130262.1)

Examination of the BLAST result

The genomic BLAST search result for accession GQ892012.1, preserved under RID [NZHNKGFY014](#), is shown to the right. The top hit in the Descriptions table is to a region on Chromosome 5 from the GRCh38.p13 Primary Assembly (A). Clicking the title of that entry changes the display to the corresponding alignment (B) for more details. A mismatch in the alignment indicates the presence of a potential variation (circled), about 1.1 kb upstream the translation start site (C).

BLAST » blastn suite » results for RID-NZHNKGFY014 [Home](#) [Recent Results](#) [Saved Strategies](#) [Help](#)

[< Edit Search](#) [Save Search](#) [Search Summary](#) [How to read this report?](#) [BLAST Help Videos](#) [Back to Traditional Results Page](#)

Job Title gb|GQ892012.1
RID NZHNKGFY014 [Search expires on 09-16 03:38 am](#) [Download All](#)
Program BLASTN [Citation](#)
Database refseq_representative_genomes (GPIPE/9606/109.20200522/ref_top_level) [See details](#)
Query ID GQ892012.1
Description Homo sapiens haplotype HHE*3 interleukin-4 (IL-4) gene, pro ...
Molecule type dna
Query Length 1340
Other reports [Distance tree of results](#) [MSA viewer](#)

Filter Results
Organism only top 20 will appear ☐ exclude
 Type common name, binomial, tid or group name
[+ Add organism](#)
Percent Identity to **E value** to **Query Coverage** to
[Filter](#) [Reset](#)

Descriptions [Graphic Summary](#) [Alignments](#) [Taxonomy](#)

Sequences producing significant alignments [Download](#) [Manage Columns](#) [Show](#) 100 [?](#)

☒ select all 100 sequences selected [GenBank](#) [Graphics](#) [Distance tree of results](#)

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
A Homo sapiens chromosome 5, GRCh38.p13 Primary Assembly	2459	81183	100%	0.0	99.78%	NC_000005.10
Homo sapiens chromosome 2						

Alignment view Pairwise ☒ CDS feature [Restore defaults](#)

100 sequences selected [Download](#) [GenBank](#) [Graphics](#) [Sort by:](#) E value [Next](#)

Homo sapiens chromosome 5, GRCh38.p13 Primary Assembly
 Sequence ID: [NC_000005.10](#) Length: 181538259 Number of Matches: 677

Range 1: 132672742 to 132674081 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Identities	Gaps	Strand
2459 bits(1331)	0.0	1337/1340(99%)	0/1340(0%)	Plus/Plus

Query 1 GCCCCAAGATCCTGCTGCCATGAGCCCTTACTGGGAGGTGGGCTTGACAGGGC 60
Sbjct 132672742 GCCCCAAGATCCTGCTGCCATGAGCCCTTACTGGGAGGTGGGCTTGACAGGGC 132672801

Query 181 TAGGAGGGCTGATTGTAAAGTTGGTAAGCAGTCTTTTCTAATTAGCTGAGGAT 1
Sbjct 132672922 TAGGAGGGCTGATTGTAAAGTTGGTAAGCAGTCTTTTCTAATTAGCTGAGGAT 1

Query 1201 TTCAGCAATTTTAAATCTATATATAGAGATATCTTTGTCAGCATTGCATCGTAGCTCT 1
Sbjct 132673942 TTCAGCAATTTTAAATCTATATATAGAGATATCTTTGTCAGCATTGCATCGTAGCTCT 1

CDS:interleukin-4, p 1 CCTGATAAACTAATTGCTCATTGTCTGCAAAATCGACACCTATTAAATGGGCTCAC 1320
Query 1261 ATTGTCAGTCAAAATCGACACCTATTAAATGGGCTCAC 132674061

Distance between 211 and 1310, ~1.1 kb

Configuration Page

Tracks [Custom Data](#)

Active Tracks [Search Tracks](#)

Category: dbSNP Build 154 (Homo sapiens Annotation Release 109) (30 Items)

Sequence ☒ Cited Variations, dbSNP b154 v2

Genes/Products ☒ Clinical, dbSNP b154 v2

Variation ☐ 1000 Genomes Phase 3, dbSNP b154 v2

Phenotype and Disease ☐ 3' Gene Region, dbSNP b154 v2

Alignments ☐ 3' UTR Region, dbSNP b154 v2

Genomic Clones ☐ 5' Gene Region, dbSNP b154 v2

Expression ☐ 5' UTR Region, dbSNP b154 v2

Epigenomics

Comparative Genomics

Features

BLAST

Track Settings: Model based Paralogous Sequence Differences [Track legend](#)

Identification of sequence differences identified from alignment of paralogous gene

Rendering options: Display on a single line

Sort variants by: No sorting

Heatmap Threshold: None

[Remove track\(s\)](#) [Configure](#) [Reset tracks](#) [Cancel](#)

Cited Variations, dbSNP b154 v2

246 | T/C rs2243247 | G/A rs2243248 | T/A/C/G rs34185442 | AAAAAAAAAAAAAAAAAA rs71645906 | C/T rs2234648 | C/T rs2243250 | C/G

Clinical, dbSNP b154 v2

Warning: No track data found in this range

5' UTR Region, dbSNP b154 v2

Genes

(U) Cleaned Alignments - BLAST Results for: gb|GQ892012.1

E

Visualizing SNP-related tracks for a genome sequence

The addition Cited Variations tracks to the graphical display (below) shows all mapped variations with association publications (A). All three mismatches in this alignment appear to correlate with annotated variations that are referenced in publications (B). Right-click a region around a desired variant to see the context menu, select the “Zoom To Sequences” option (C) to see more details at the sequence level.

The screenshot shows the NCBI genome browser interface. The top track displays the sequence alignment for NC_000005.10. Below it, the 'Cited Variations, dbSNP b154 v2' track shows various SNPs. A context menu is open over a mismatched position (G vs T), with options like 'Zoom To Sequence' highlighted. Red dashed lines connect the menu options to the corresponding views in the subsequent screenshots.

The screenshot shows a zoomed-in view of the sequence alignment. The mismatched position (G vs T) is clearly visible. A popup box (E) shows details for the variation rs2243248, including its ID, type, alleles, and genomic locations.

In the sequence-level view (D), we can see that the specific change the input query has at this mismatched position is G, vs the T in the reference genomic sequence.

Hovering over a variation activates a popup box with additional details about the variation (E). Clicking the rsID opens the SNP records in a new window (F). The PubMed link (G) points to the set of publications with research findings on this specific SNP record.

The screenshot shows the dbSNP Short Genetic Variations 'Reference SNP (rs) Report' for rs2243248. The report includes details about the organism (Homo sapiens), position (chr5:132672952), alleles (T>A / T>C / T>G), variation type (SNV Single Nucleotide Variation), and frequency. It also lists clinical significance and publications.

The screenshot shows the PubMed.gov search results for the SNP rs2243248. The search results show 54 results, with the top two results displayed. The first result is 'Associations between Interleukin Gene Polymorphisms and Risks of Developing Extremity Posttraumatic Osteomyelitis in Chinese Han Population.' The second result is 'Association of Interleukin-4 Polymorphisms With Breast Cancer in Taiwan.'

SNP mapping to RefSeq mRNAs

The RefSeq RNA BLAST result for CK130262.1, preserved under RID [P09WMZYP01R](#), is shown below and to the right. The top hit in the Descriptions table is the IL4 transcript variant NM_000589.4 (A). Clicking the record title change the page to display the alignment (B) where three mismatches (circled) are shown. Clicking the Graphics link (C) displays the alignment in SV to allow interactive examination of variations in the query under the context of NCBI annotations.

Descriptions	Graphic Summary	Alignments	Taxonomy
--------------	-----------------	------------	----------

Sequences producing significant alignments						
Download Manage Columns Show 100						
select all 4 sequences selected						
GenBank Graphics Distance tree of results						
Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
A Homo sapiens interleukin 4 (IL4), transcript variant 1, mRNA	841	841	78%	0.0	95.98%	NM_000589.4
Homo sapiens interleukin 4 (IL4), transcript variant 3, mRNA	436	846	78%	2e-120	98.78%	NM_001354990.2
Homo sapiens interleukin 4 (IL4), transcript variant 2, mRNA						
Homo sapiens uncharacterized LOC105379176 (LOC105379176), long non-coding RNA						

Customizing the SV display around the first mismatch by zooming to the sequence level using context menu option (D) and adding the Cited Variation track under Variation (E), it is clear that the first mismatch corresponds with the existing variation record rs2070874 (F). To better identify the variations using the naming convention from the Human Genome Variation Society (HGVS), you can locate the position of A for the start codon, right-click and select "Set Sequence Origin At Position" (G) to reset the numbering of this mRNA to begin at the start of the open reading frame. The adjusted position of the variant is 33 bases upstream from the start of the coding start.

Homo sapiens interleukin 4 (IL4), transcript variant 1, mRNA
Sequence ID: [NM_000589.4](#) Length: 615 Number of Matches: 1

Range 1: 1 to 514 [GenBank](#) [Graphics](#) [Next Match](#)

Score	Expect	Identities	Gaps	Strand
841 bits(455)	0.0	502/523(96%)	10/523(1%)	Plus/Plus

Query 38 ATCGTTAGCTTCTCCTGATAAACTAAATGTCACACATTGCTCAATCGACACCTAT 97
Sbjct 1 ATCGTTAGCTTCTCCTGATAAACTAAATGTCACACATTGCTCAATCGACACCTAT 60

Query 98 TAATAGGCTCACCTCCCAACTACCTCCCTCTGTTCTTCTGCTAGCATGTGCCGGCA 157
Sbjct 61 TAATAGGCTCACCTCCCAACTGCTCCCTCTGTTCTTCTGCTAGCATGTGCCGGCA 120

Query 158 ACTTTGTCTCACGGACACAAAGTGCATATCACCTTACAGGAGATCATCAAACTTTGAACA 217
Sbjct 121 ACTTTGTCTCACGGACACAAAGTGCATATCACCTTACAGGAGATCATCAAACTTTGAACA 180

Configure Page

Tracks Custom Data

Active Tracks

Active Track name

☒ Cited Variations, dbSNP b154 v2

☐ Clinical, dbSNP b154 v2

☐ 1000 Genomes Phase 3, dbSNP b154 v2

☐ 3' UTR Region, dbSNP b154 v2

☐ 5' UTR Region, dbSNP b154 v2

☐ Common Variations (MAF >= 0.01), dbSNP b154 v2

☐ ExAC Release 1 Frequency, dbSNP b154 v2

Track Settings: Cited Variations, dbSNP b154 v2 [Track legend](#)

dbSNP 2.0 Build 154 v2 all data based on Homo sapiens

Rendering options: Show variants for 50 or less

[Remove track\(s\)](#) [Configure](#) [Reset tracks](#) [Cancel](#)

Marker Details

Name: IL4.c-33C>T Position: -33

Accession/Locus tag Location Relative to HGVS Name Sequence

NM_000589.4	30	Seq start	NM_000589.4:r.30	AACTAATTGCTCATTGT
NM_000589.4	-33	Current origin	NM_000589.4:r.30	AACTAATTGCTCATTGT

[Download](#) [Close](#)

References

1. The Graphical Sequence Viewer: ftp.ncbi.nih.gov/pub/factsheets/Factsheet_Graphical_SV.pdf
2. The New BLAST Results Page: ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_NewResultPage.pdf
3. SNP: Database for Short Genetic Variations: ftp.ncbi.nlm.nih.gov/pub/factsheets/Factsheet_SNP.pdf

The "Marker Details" option in the marker context menu (not shown) displays a table (H) to provide the location relative to the newly set sequence origin. This takes manual counting out of the process, and the information put the HGVS annotation of this variant as NM_000589.4:c.-33C>T.